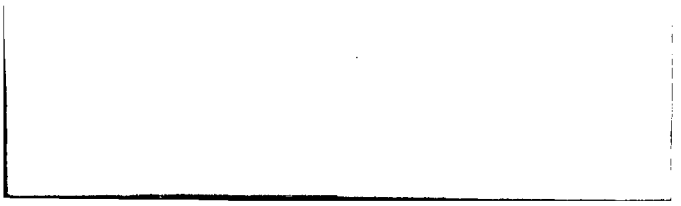
AD-A118 245 DESMATICS INC STATE COLLEGE PA F/G 5/10
DATA INTEGRATION: COMBINING REAL-WORLD AND SIMULATION DATA.(U)
AUG 82 D E SMITH N00014-75-C-1054
UNCLASSIFIED TR-106-12 NL

END
DATE FILMED
9 82
DTIC



DESMATICS, INC.

P. O. Box 618
State College, Pa. 16801
Phone: (814) 238-9621

Applied Research in Statistics - Mathematics - Operations Research

DATA INTEGRATION: COMBINING REAL-WORLD AND SIMULATION DATA

by

Dennis E. Smith

TECHNICAL REPORT NO. 106-12

August 1982

This study was supported by the Office of Naval Research under Contract No. N00014-75-C-1054, Task No. NR 042-334 and Contract No. N00014-79-C-0650, Task No. NR 277-291

TABLE OF CONTENTS

I. INTRODUCTION

In any form of scientific research or decision making, it is desirable to draw upon all relevant data which is available. Unfortunately, data derived from different sources often takes on forms which are incompatible. Consequently, much of the information is often not used and is thereby effectively "lost."

Simulation users frequently find themselves in this situation when observations have been obtained both from a computer model and from the corresponding real-world situation it simulates. Although the real-world observations comprise the most valid of the two data sets, the other set may also contain useful information.

In general, real-world (experimental) observations are subject to statistical variation. Simulation outputs from a model of the same situation not only contain statistical variation, but also may be clouded by possible model inadequacies. If the model is a valid representation of the corresponding real-world situation, then the two types of data (simulation and experimental) are of equal value. However, if model validity is in question, the simulation data may be of less value than the experimental data. This raises the issue of model validation, which has been discussed by a number of authors (e.g., [1], [2], [3], [4], [5]). The validation task is to compare the simulation data with that of the real-world system, usually by means of a hypothesis test.

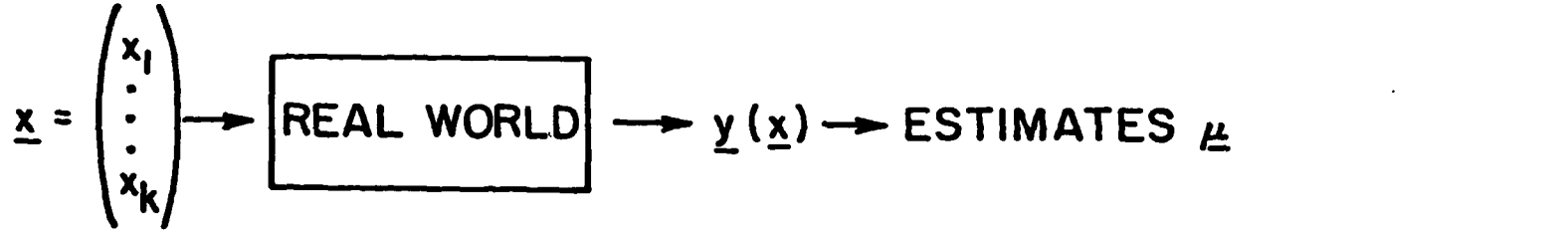
This paper does not deal with validation, per se, but rather with a topic we label "data integration." With data integration, we aren't interested in a yes/no decision about whether or not the simulation is valid. Instead, we are concerned with whether the simulation data is useful. In

other words, we want to determine how we can best use the simulation data to supplement the real-world observations. Thus, the focus of data integration is to determine the best procedure for combining data obtained from a simulation and from the corresponding real-world situation.

In this paper, we assume that we are dealing with situations in which one simulation run yields a single response vector, rather than a time series vector. Thus, we are restricting our attention to terminating simulations. Figure 1 illustrates this framework, in which we have two data sources, each of which generates a response vector which is a function of $\underline{x}$, a vector of input variables.

For purposes of this paper, we assume that each observation is produced under identical conditions, i.e., for a specific value $\underline{x} = \underline{x}_0$. In view of this, we will suppress the dependence of the response on $\underline{x}$ in the ensuing discussion. The real-world response $\underline{y}$ estimates some unknown parameter vector $\underline{\mu}$, contaminated by random error. The simulation response $\underline{w}$ which is supposed to estimate $\underline{\mu}$ is not only affected by random error, but also may be biased because of inaccuracies in the simulation model. Our data integration goal is to obtain the most accurate estimate of $\underline{\mu}$ that we can.

- TWO DATA SOURCES:



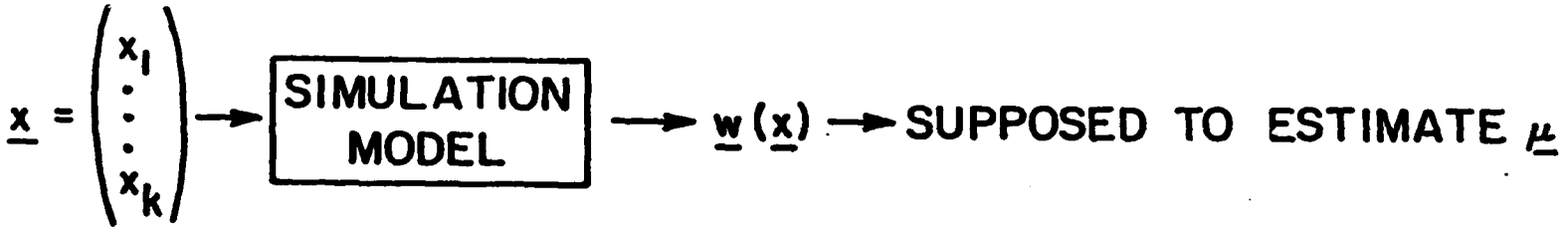


Figure 1: Illustration of the Problem Framework

II. PROBLEM DISCUSSION

In this paper we examine the univariate case. Specifically, consider the situation in which a sample of n independent real-world observations $y_1, \ldots, y_n$ is observed, where $y_i \sim N(\mu,\sigma^2)$ and the object is to estimate μ. Assume that in addition to, and independent of, the y_i's, m independent observations $w_1, \ldots, w_m$ are available from a simulation model for which $w_j \sim N(\mu+\Delta\sigma, \sigma^2)$. Thus, if $\Delta \neq 0$, the simulation data contains a bias. As an aside, we should note that, in general, $n < m$ and n is usually quite small because of the difficulty and/or expense of obtaining real-world observations.

Suppose we decide to estimate μ by using an estimator of the form

$$\hat{\mu}_p = p\bar{y} + (1-p)\,\bar{w} ,$$

which is a weighted average of the real-world responses and the simulation responses. The pooled mean, in which each observation is weighted equally, is obtained when $p = p^* = n/(n+m)$, but this estimator would be optimal only if the simulation model were valid, i.e., if $\Delta = 0$. Since the assumption of a simulation model being valid is at best tenuous, we will examine what happens for values of $\Delta \neq 0$, adopting mean square error (MSE) as the measure of the goodness of an estimator.

For general p, where p is a constant,

$$MSE(\hat{\mu}_p) = p^2\sigma^2/n + (1-p)^2\,[(\sigma^2/m)+\Delta^2\sigma^2] \qquad (1)$$

If we chose always to use only the real-world data, then our estimator would be $\hat{\mu}_1$, which is nothing more than $\bar{y}$. For this estimator

$$MSE(\hat{\mu}_1) = \sigma^2/n . \qquad (2)$$

On the other hand, if we chose always to use both data sources, giving each observation equal weight, our estimator would be the pooled mean

$$\hat{\mu}_p^* = (n\bar{y} + m\bar{w})/(n+m) \ .$$

The corresponding MSE is

$$MSE(\hat{\mu}_{p*}) = \sigma^2(n+m+m^2\Delta^2)/(n+m)^2 \qquad (3)$$

As one would expect, for values of Δ close to zero, the estimator $\hat{\mu}_p^*$ based on both sets of observations would provide a smaller MSE than that resulting from the use of the real-world data alone. In fact, we can see from equations (2) and (3) that the use of $\hat{\mu}_p^*$ provides better performance (i.e., smaller MSE) so long as $|\Delta| < [(n+m)/nm]^{\frac{1}{2}}$. However, for larger values of $|\Delta|$, the inflation in MSE rapidly becomes catastrophic; the MSE is unbounded as $|\Delta| \to \infty$.

Of course, we could avoid such catastrophic results by never using the simulation observations, i.e., by always using the estimator $\hat{\mu}_1 = \bar{y}$. However, by adopting this conservative minimax strategy, we would deprive ourselves of the opportunity to obtain much better estimates when Δ is small.

III. TWO DATA INTEGRATION APPROACHES

One approach to this problem is to return to the validation framework, and use the estimator $\hat{\mu}_{p*}$ if the simulation model is judged valid or the estimator $\hat{\mu}_1$, otherwise. As mentioned previously, a judgement about the validity of a simulation model is usually based on a hypothesis test. In the situation being discussed, the appropriate hypothesis test would involve the hypotheses

$$H_0: \Delta = 0$$

$$H_1: \Delta \neq 0$$

based on the t-statistic

$$t=[nm/(n+m)]^{\frac{1}{2}}(\bar{y}-\bar{w})/s$$

where s denotes the pooled estimated standard deviation. The rejection region (assuming a significance level of α) would be

$$|t|>t_{\alpha/2,n+m-2}$$

where $t_{\alpha/2,\ n+m-2}$ denotes the upper $\alpha/2$ point of a t-distribution with $n+m-2$ degrees of freedom.

If H_0 were rejected, the estimator $\hat{\mu}_1 = \bar{y}$ would be used, while if it were not rejected, the estimator $\hat{\mu}_{p*} = (n\bar{y}+m\bar{w})/(n+m)$ would be used. It should be noted that an estimate of Δ is given by

$$\hat{\Delta} = (\bar{y}-\bar{w})/s \ .$$

Therefore, the validation approach results in the use of the estimator

$$\hat{\mu} = \begin{cases} \hat{\mu}_{p*} , & \text{if } |\hat{\Delta}|<[(n+m)/nm]^{\frac{1}{2}}t_{\alpha/2,n+m-2} \\ \hat{\mu}_1 , & \text{if } |\hat{\Delta}|\geq[(n+m)/nm]^{\frac{1}{2}}t_{\alpha/2,n+m-2} \end{cases}$$

where, of course, the value of α must be specified. This estimator, it will be noted, is based on an all or nothing rule---if $|\hat{\Delta}|$ is too large, only the real-world observations are used, whereas if $|\hat{\Delta}|$ is not too large, all observations are used. Thus, a simulation observation is given weight zero or equal weight with each real-world observation, depending upon the size of $\hat{\Delta}$.

A more flexible procedure would incorporate $\hat{\Delta}$ directly into the estimate. Suppose, therefore, in view of the fact we are unwilling to accept the assumption of $\Delta = 0$, we attempt to determine an adaptive method for incorporating information about Δ into the estimator of μ. Using equation (1), we can see that by setting

$$\partial \mathrm{MSE}(\hat{\mu}_p)/\partial p = 0 \ ,$$

we find that

$$p = (n+nm\Delta^2)/(n+m+nm\Delta^2) \tag{4}$$

provides the minimum MSE. It should be noted that if $\Delta = 0$, $p = n/(n+m)$ so that $\hat{\mu}_p$ reduces to the weighted average which we would use if it were assumed that the simulation observations should be given the same weight as the real-world observations.

Of course, because Δ is an unknown parameter, the value of p providing the minimum MSE is also unknown. Thus, we might consider substituting $\hat{\Delta}$ into (4) and using the resulting value

$$\hat{p} = (n+nm\hat{\Delta}^2)/(n+m+nm\hat{\Delta}^2) \ .$$

This results in an adaptive estimator $\hat{\mu}_{\hat{p}}$. Because $\hat{p}$ is a random variable rather than a constant, $\mathrm{MSE}(\hat{\mu}_{\hat{p}})$ cannot be obtained by substitution into equation (1).

IV. EVALUATION OF MSE

At this juncture, we have four estimators to consider in examining the data integration task. These are:

(a) $\hat{\mu}_{p*}$, which always uses the real-world and simulation observations weighted equally,

(b) $\hat{\mu}_1$, which always uses only the real-world observations,

(c) $\hat{\mu}$, which is based on a test of the validity of the simulation model,

and (d) $\hat{\mu}_{\hat{p}}$, which is an adaptive estimator.

We note that an investigation of these estimators does not depend on the actual values of μ and σ, since location has no effect on the results and the bias of any simulation observation is measured in units of σ.

In order to compare the performance of these four estimators for a sample size (n,m), their MSE's must be evaluated for different values of Δ. This poses no difficulty in the case of the first two estimators ($\hat{\mu}_{p*}$ and $\hat{\mu}_1$); the required MSE's are given by equations (1) and (2). Unfortunately, things aren't so easy when considering the estimators $\hat{\mu}$ and $\hat{\mu}_{\hat{p}}$. For $\hat{\mu}$, we must compute the expected value of

$$[(n\bar{y} + m\bar{w})/(n+m) - \mu]^2$$

over the region in $(s, \bar{y}, \bar{w})$ - space

$$|(\bar{y} - \bar{w})/s| < [(n+m)/nm]^{\frac{1}{2}} t_{\alpha/2, n+m-2},$$

and the expected value of $(\bar{y} - \mu)^2$ over the region

$$|(\bar{y} - \bar{w})/s| \geq [(n+m)/nm]^{\frac{1}{2}} t_{\alpha/2, n+m-2}.$$

For the adaptive estimator $\hat{\mu}_{\hat{p}}$, we must evaluate the expected value of

$$[\{n + nm[(\bar{y} - \bar{w})/s]^2\}\bar{y} + m\bar{w}]/[\{n + m + nm[(\bar{y} - \bar{w})/s]^2\} - \mu]^2 .$$

Because of their complexity, an analytic evaluation of these expected values is an impossible task. Thus, we must turn to numerical integration or to Monte Carlo. Since we could not eliminate the need to evaluate triple integrals, we chose to use Monte Carlo to investigate the two specific cases of (n=3, m=10) and (n=3, m=50).

Tables 1 and 2 list the MSE of $\hat{\mu}_{p^*}$ (the pooled mean estimator), of $\hat{\mu}$ (the validity test estimator) and of $\hat{\mu}_{\hat{p}}$ (the adaptive estimator) relative to that of $\hat{\mu}_1 = \bar{y}$, which is $\sigma^2/3$ in both cases. For the validity test estimator, five values of α were considered. These were .01, .05, .10, .20, and α^*, where α^* denotes the value of α which provides an MSE equal to that of the adaptive estimator $\hat{\mu}_{\hat{p}}$ when $\Delta = 0$. For (n=3, n=10), $\alpha^* = .17$ which corresponds to a t value of 1.50, while for (n=3, m=50), $\alpha^* = .12$ which corresponds to a t value of 1.58.

$\|\Delta\|$	Pooled Mean Estimator $\hat{\mu}_{p^*}$	Validity Test Estimator $\hat{\mu}$					Adaptive Estimator $\hat{\mu}_{\hat{p}}$
		α=.01	α=.05	α=.10	α=.20	$\alpha=\alpha^*$	
0.00	0.23	0.29	0.39	0.50	0.66	0.62	0.62
0.25	0.34	0.39	0.52	0.57	0.73	0.73	0.63
0.50	0.68	0.77	0.91	0.97	0.99	1.01	0.75
0.75	1.23	1.34	1.38	1.34	1.22	1.27	0.88
1.00	2.01	2.02	1.75	1.58	1.32	1.52	1.07
1.50	4.22	3.34	2.27	1.82	1.38	1.49	1.20
2.00	7.33	3.96	2.19	1.60	1.25	1.33	1.17
3.00	16.21	2.80	1.32	1.17	1.13	1.14	1.11
4.00	28.63	1.28	1.15	1.13	1.10	1.11	1.05
5.00	44.61	1.09	1.07	1.05	1.04	1.05	1.04

Table 1: MSE's of estimators relative to $MSE(\hat{\mu}_1) = MSE(\bar{y})$ for $n = 3$, $m = 10$
(MSE's for $\hat{\mu}$ and $\hat{\mu}_{\hat{p}}$ estimated by Monte Carlo; maximum standard error is 0.04)

$\lvert\Delta\rvert$	Pooled Mean Estimator $\hat{\mu}_p$*	Validity Test Estimator $\hat{\mu}$					Adaptive Estimator $\hat{\mu}_{\hat{p}}$
		α=.01	α=.05	α=.10	α=.20	α=α*	
0.00	0.06	0.12	0.27	0.43	0.64	0.48	0.48
0.25	0.22	0.32	0.49	0.61	0.76	0.72	0.52
0.50	0.72	0.88	0.98	1.04	1.06	1.03	0.73
0.75	1.56	1.69	1.64	1.55	1.37	1.53	1.02
1.00	2.73	2.63	2.16	1.87	1.54	1.76	1.11
1.50	6.06	3.91	2.53	1.95	1.44	1.62	1.20
2.00	10.74	3.08	1.70	1.31	1.18	1.28	1.26
3.00	24.09	1.00	1.00	1.01	1.17	1.04	1.07
4.00	42.78	0.99	0.99	0.99	1.00	1.00	1.05
5.00	66.81	0.97	0.98	0.98	1.01	0.99	1.03

Table 2: MSE's of estimators relative to $MSE(\hat{\mu}_1) = MSE(\bar{y})$ for $n = 3$, $m = 50$
(MSE's for $\hat{\mu}$ and $\hat{\mu}_{\hat{p}}$ estimated by Monte Carlo; maximum standard error is 0.05)

V. DISCUSSION

As can be seen from Tables 1 and 2, and graphically from Figures 2 and 3, none of the four estimators dominates (or is dominated by) any other estimator in terms of MSE. This in itself is not surprising. What is surprising, and somewhat disconcerting, is that the simulation data is useful (i.e., provides a more accurate estimate of μ) only if Δ is very small. For no matter which estimator (other than $\bar{y}$) we adopt, we can never come out ahead if $|\Delta|>\sigma$, and in fact we may wind up doing substantially worse than we may have thought possible.

It is clear that the pooled mean estimator $\hat{\mu}_{p^*}$, with its unbounded MSE is not worth considering. By using either the validity test estimator $\hat{\mu}$ or the adaptive estimator $\hat{\mu}_{\hat{p}}$, we will come out ahead, or at least not too far behind, if $|\Delta|$ is either small or large. It is for moderate values of $|\Delta|$, approximately $1.0 <|\Delta|<3.0$, that the worst things happen to us. Therefore, somewhat with tongue in cheek, we see that the resulting moral is to construct either a very accurate simulation or a very inaccurate one.

More seriously, though, our results indicate that a test of validity, per se, is unwarranted (and hazardous) if the data is to be used for parameter estimation. We can see this from Figures 2 and 3 by examining the results of a validity test at the usual significance levels of .01 and .05. If we wish to take a chance on combining real-world and simulation data, $\hat{\mu}_{\hat{p}}$ appears to be our best bet since it provides reasonable gains (decreases in MSE) for small $|\Delta|$ and in the worse case does not substantially penalize us.

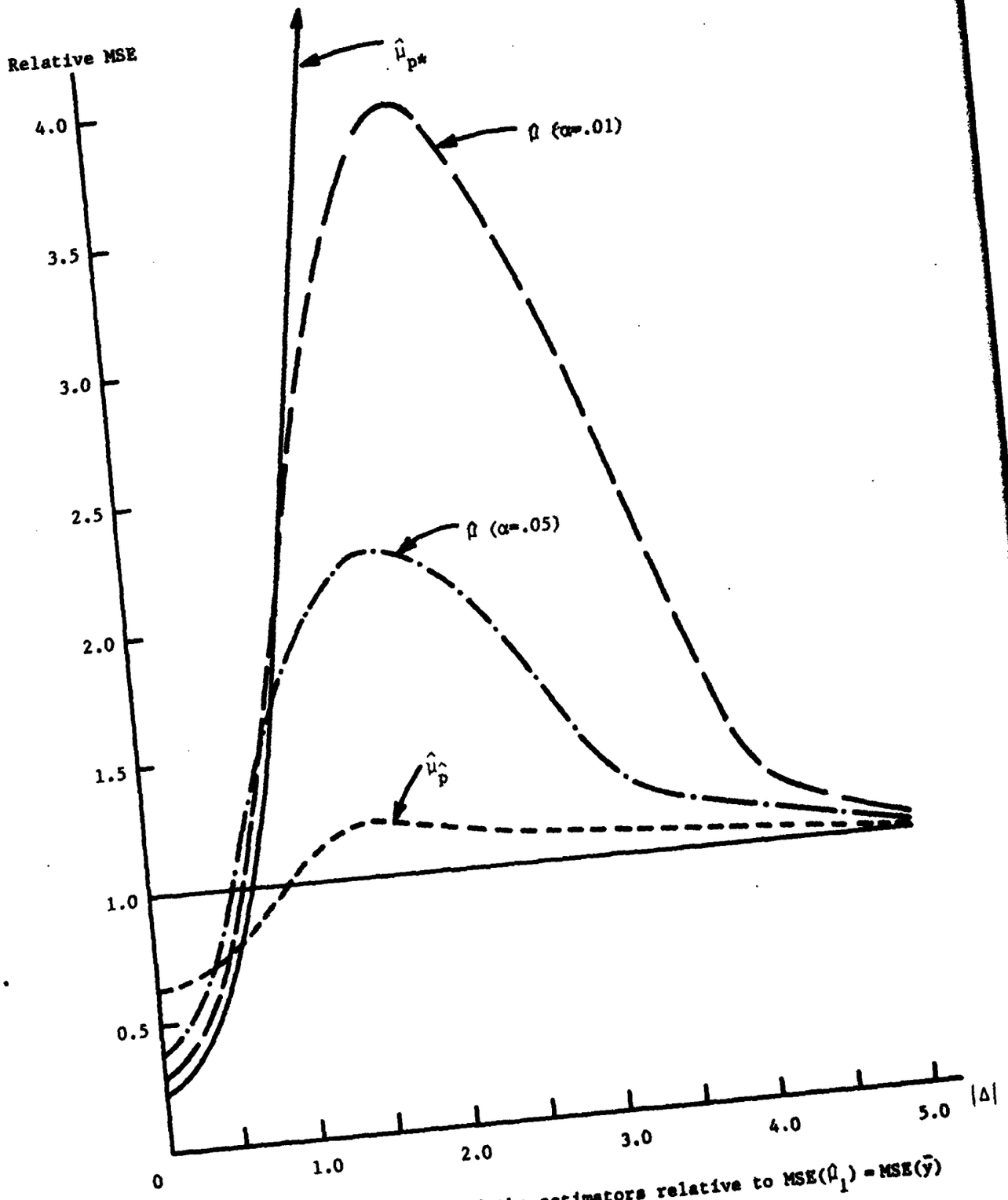


Figure 2: MSE's of the estimators relative to $MSE(\hat{\mu}_1) = MSE(\bar{y})$ for $n = 3$, $m = 10$

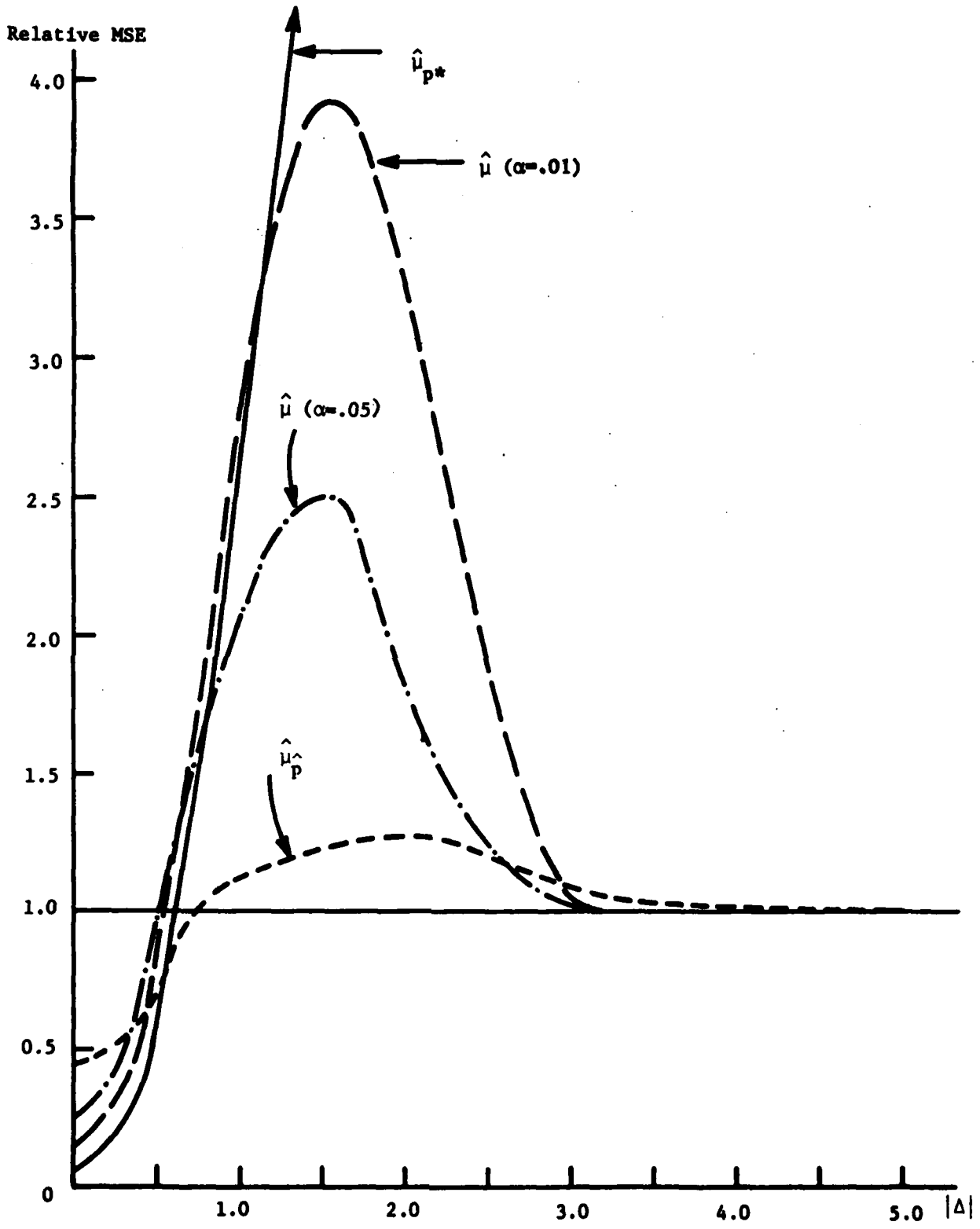


Figure 3: MSE's of the estimators relative to $MSE(\hat{\mu}_1) = MSE(\bar{y})$ for $n = 3$, $m = 50$

VI. REFERENCES

[1] Law, A. M. and Kelton, W. D., Simulation Modeling and Analysis, McGraw-Hill Book Co., New York, 1982.

[2] Naylor, T. H. and Finger, J. M., "Verification of Computer Simulation Models," Management Science, Vol. 14, pp. 92-101, 1967.

[3] Sargent, R. G. "Validation of Simulation Models," Proceedings of the Winter Simulation Conference, pp. 497-503, 1979.

[4] Schellenberger, R. E., "Criteria for Assessing Model Validity for Managerial Purposes," Decision Science, Vol. 5, pp. 644-653, 1974.

[5] Van Horn, R. L., "Validation of Simulation Results," Management Science, Vol. 17, pp. 247-258, 1971.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 106-12	2. GOVT ACCESSION NO. AD-A118245	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) DATA INTEGRATION: COMBINING REAL-WORLD AND SIMULATION DATA		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Dennis E. Smith		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-1054 N00014-79-C-0650
9. PERFORMING ORGANIZATION NAME AND ADDRESS Desmatics, Inc. P. O. Box 618 State College, PA 16801		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 042-334 NR 277-291
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Arlington, VA 22217		12. REPORT DATE August 1982
		13. NUMBER OF PAGES 15
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this report is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Data Integration
Simulation Validation
Simulation Data Analysis
Model Validation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In this paper we discuss a topic we label "data integration," which addresses the problem of combining information from a simulation model with that from the corresponding real-world situation. Although data integration is related to simulation validation, it does not focus on a yes/no decision about whether or not a model is valid. Instead, it is concerned with whether or not the simulation data is useful.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)